Classifying Matrices with a Spectral Regularization

Ryota Tomioka & Kazuyuki Aihara

University of Tokyo / Fraunhofer FIRST

2007/06/23

THE UNIVERSITY OF TOKYO

FIRST
Fraunhofer Institut
Rechnerarchitektur
und Softwaretechnik

BBCI
berlin
brain computer
interface

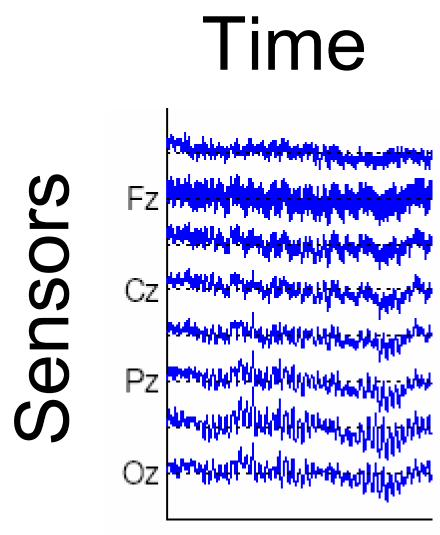
Outline

- Method
  - Discriminative model that factorizes using the spectral ℓ_1-regularization.
  - Penalized empirical loss minimization (convex!).
- Implementation
  - Dual formulation.
  - Linear Matrix Inequality.
  - Interior point method.
- Application
  - Motor-imagery EEG classification.
- Summary

Examples of Matrix Inputs

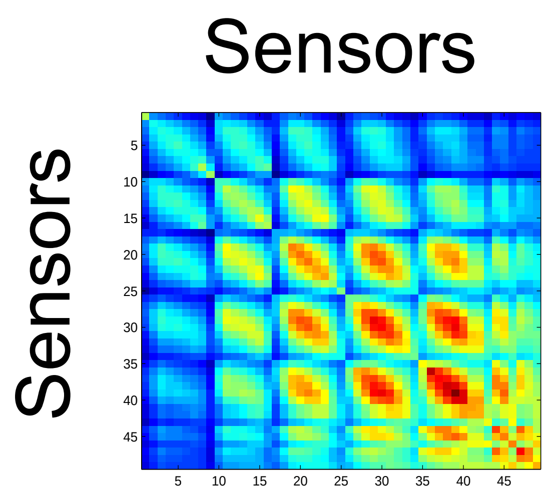
- Multivariate Time Series

$$X =$$



- Second order statistics

$$X =$$



Problem Setting

The Input

$$X$$

$R \times C$

Class Label

$$y \in \{+1, -1\}$$

$$f(X; W, b) = \mathrm{Tr}\left[W^\top X\right] + b$$

$$(W \in \mathbb{R}^{R \times C},\ b \in \mathbb{R})$$

Spectral ℓ_1-regularization (sum of singular-values):

$$\Omega(W) = \sum_{c=1}^{r} \sigma_c[W]$$

Interpreting the Model

Using the singular-value decomposition:

$$W = U \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{pmatrix} V^\top = \sum_{c=1}^{r} \sigma_c \boldsymbol{u}_c \boldsymbol{v}_c^\top$$

The classifier can be written as:

$$f(X) = \mathrm{Tr}\left[\left(\sum_c \sigma_c \boldsymbol{u}_c \boldsymbol{v}_c^\top \right)^\top X \right] + b$$

$$= \sum_{c=1}^{r} \sigma_c \boldsymbol{u}_c^\top X \boldsymbol{v}_c + b$$

Interpreting the Model

Using the singular-value decomposition:

$$W = U \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{pmatrix} V^\top = \sum_{c=1}^{r} \sigma_c \boldsymbol{u}_c \boldsymbol{v}_c^\top$$

The classifier can be written as:

$$f(X) = \mathrm{Tr}\left[\left(\sum_c \sigma_c \boldsymbol{u}_c \boldsymbol{v}_c^\top \right)^\top X \right] + b$$

$$= \sum_{c=1}^{r} \sigma_c \boldsymbol{u}_c^\top X \boldsymbol{v}_c + b$$

Linear combination

Features (projected inputs)

Comments on Related Methods

LASSO:

$$\Omega_{LASSO}(W) = \sum_{(i,j)} \left| W_{ij} \right|$$

Ridge penalty:

$$\Omega_2(W) = \frac{1}{2} \sum_{i,j} W_{ij}^2 = \sum_{c=1}^{r} \sigma_c^2 [W]$$

Spectral ℓ_1-regularization:

$$\Omega_1(W) = \sum_{c=1}^{r} \sigma_c [W]$$

The Problem

(P)
$$\min_{W\in\mathbb{R}^{R\times C},\, b\in\mathbb{R},\, \boldsymbol{z}\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^{n}\ell_{LR}(z_i) + \frac{\lambda}{n}\|W\|_1,$$

s.t.
$$y_i\left(\mathrm{Tr}\left[W^\top X_i\right] + b\right) = z_i \quad (i = 1, \ldots, n),$$

Lagrange multipliers (α_i)

$$\ell_{LR}(z) := \log\left(1 + \exp(-z)\right),$$

$$\|W\|_1 := \sum_{c=1}^{r}\sigma_c[W]$$

Implementation

- Dual Formulation
- Linear Matrix Inequality
- Interior Point Method

The First Trick: The Dual Optimization Problem

(D)

$$\min_{0 \leq \boldsymbol{\alpha} \leq 1} \sum_{i=1}^{n} \ell^{*}_{\mathrm{LR}}(\alpha_i)$$

The fit must be simple (large entropy)

$$\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0,$$

$$\left\| \sum_{i=1}^{n} \alpha_i y_i X_i \right\|_{\infty} \leq \lambda,$$

Residual of the fit must be small

ℓ_∞-norm

$$\ell^{*}_{\mathrm{LR}}(\alpha) := \alpha \log \alpha + (1-\alpha) \log (1-\alpha),$$

$$\|X\|_{\infty} := \max_{c} \sigma_c [X].$$

The Second Trick: Using Linear Matrix Inequality

$$\|A(\boldsymbol{\alpha})\|_\infty = \max \sigma \left[A(\boldsymbol{\alpha}) \right] \leq \lambda$$

$$A(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i y_i X_i$$

The Second Trick: Using Linear Matrix Inequality

$$\|A(\boldsymbol{\alpha})\|_\infty = \max \sigma\left[A(\boldsymbol{\alpha})\right] \leq \lambda$$

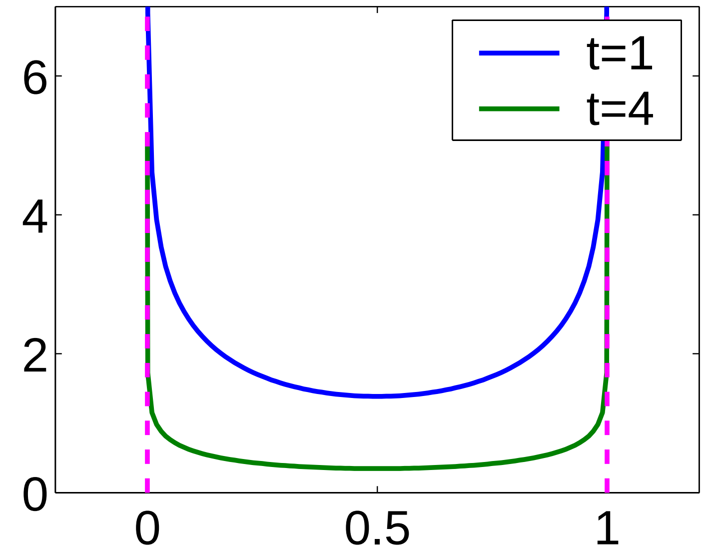
$$A(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i y_i X_i$$

$$\Leftrightarrow \begin{bmatrix} \lambda I_R & A(\boldsymbol{\alpha}) \\ A^\top(\boldsymbol{\alpha}) & \lambda I_C \end{bmatrix} \succeq 0$$

The Third Trick: Interior Point Method

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \quad \sum_{i=1}^{n} \ell^*_{\mathrm{LR}}(\alpha_i) + \frac{1}{t}\phi(\boldsymbol{\alpha}),$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0.$$



$$t \to \infty$$

Original problem!

$$\phi(\boldsymbol{\alpha}) := -\left(\log\det \begin{bmatrix} \frac{\lambda}{n} I & A(\boldsymbol{\alpha}) \\ A^\top(\boldsymbol{\alpha}) & \frac{\lambda}{n} I \end{bmatrix} + \log\alpha + \log(1-\alpha) \right).$$

$$\left(A(\boldsymbol{\alpha}) = \textstyle\sum_i \alpha_i y_i X_i\right)$$

Good News for IP optimization

- Obtaining the Primal Variable:

$\hat{\boldsymbol{\alpha}}_t$: solution at barrier parameter *t*

$$W_{\hat{\alpha}_t} = U_{\hat{\alpha}_t} \mathrm{diag}\left(\frac{2\lambda_c^{(\hat{\alpha}_t)}}{t\left(\lambda^2 - \lambda_c^{(\hat{\alpha}_t)^2}\right)} \right) V_{\hat{\alpha}_t}^{\top}$$

$$\left(U_{\hat{\alpha}_t} \Lambda_{\hat{\alpha}_t} V_{\hat{\alpha}_t}^{\top} := A(\hat{\boldsymbol{\alpha}}_t) = \sum_{i=1}^{n} \hat{\alpha}_{t,i} y_i X_i \right)$$

- Quality guarantee:

$$\mathrm{Duality\ gap}(W_{\hat{\alpha}_t}, \hat{\boldsymbol{\alpha}}_t) \leq \frac{R + C + 2n}{t}$$

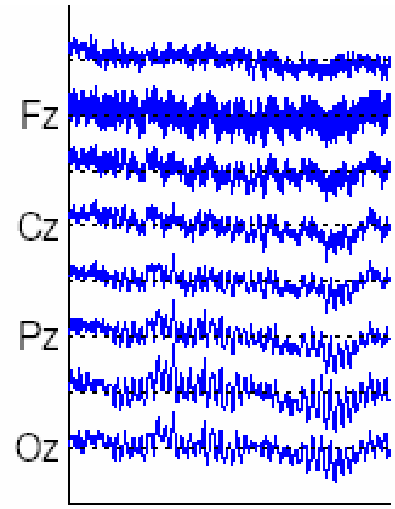
Application: Motor-imagery EEG Classification

Single-trial EEG Classification

The Covariance EEG signal Class Label

$$\underset{C \times C}{X} = SS^{\top} \Rightarrow y \in \{+1, -1\}$$

$$\underset{C \times T}{S} =$$



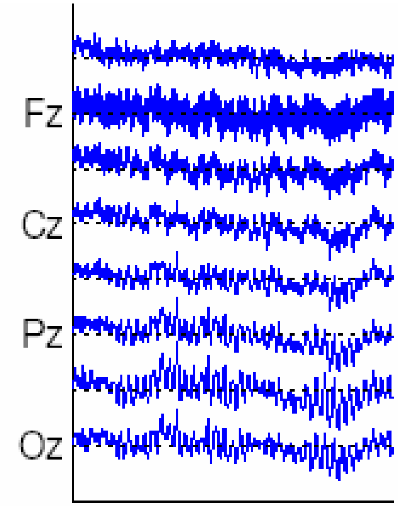
Single-trial EEG Classification

The Covariance EEG signal

$$\underset{C \times C}{X} = SS^{\top}$$

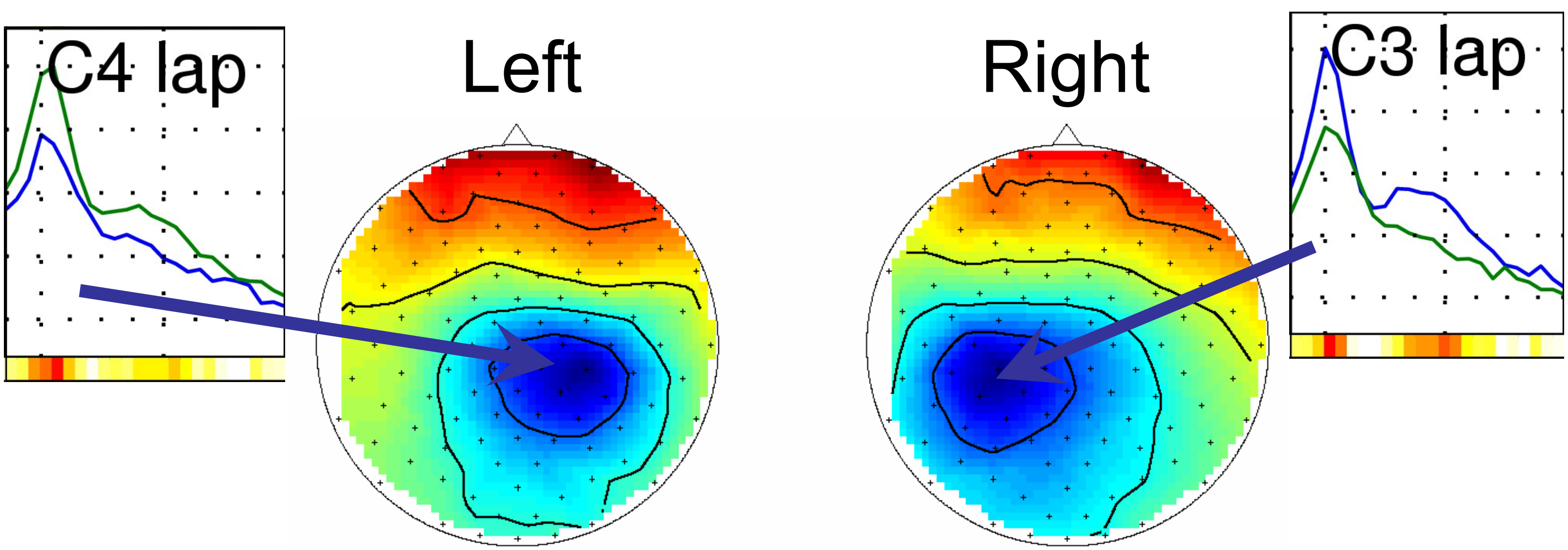
Class Label

$$y \in \{+1, -1\}$$



ERD/ERS

Lateralized modulation of rhythmic activity

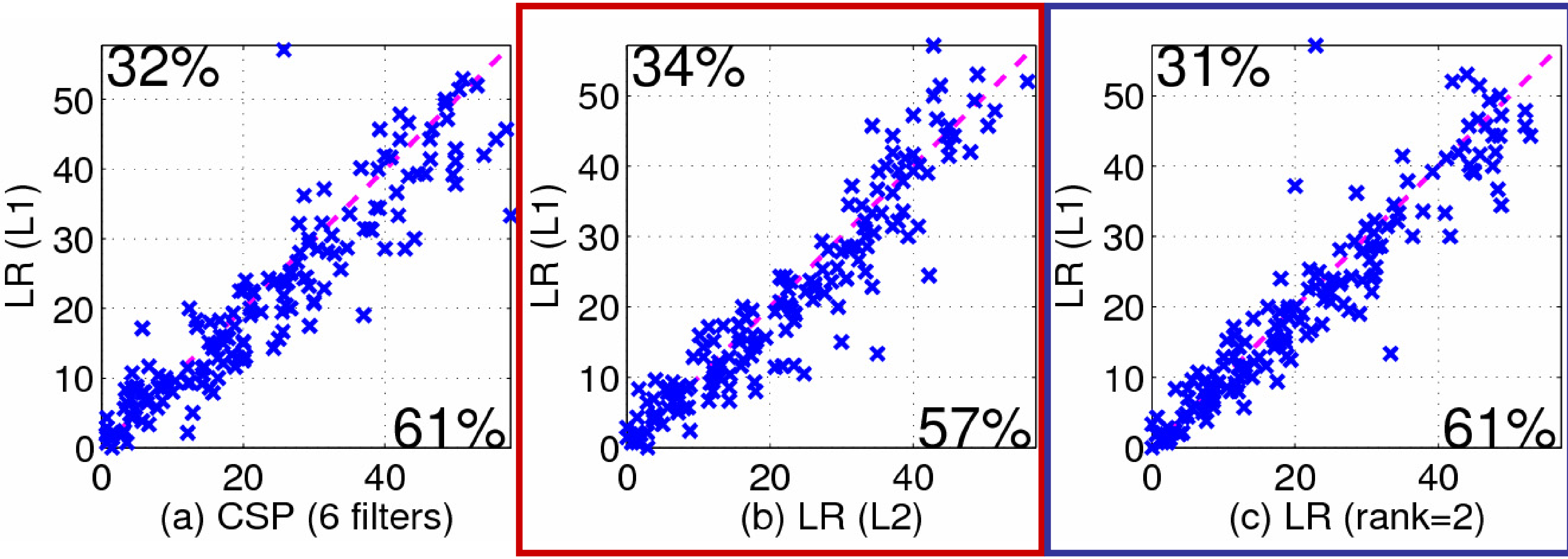


Conventional Methods

- Common Spatial Pattern (CSP) [Koles 1991; Ramoser 2000] (State of the art)
  - Two steps:
    - Feature Extraction: Find a low-dimensional decomposition.
    - Classify: linear classifier on the log-power feature.
- LR (L2)
  - ℓ_2(Frobenius norm)-regularized logistic regression.
- LR (rank=2)
  - Rank=2 constrained logistic regression (nonconvex!)

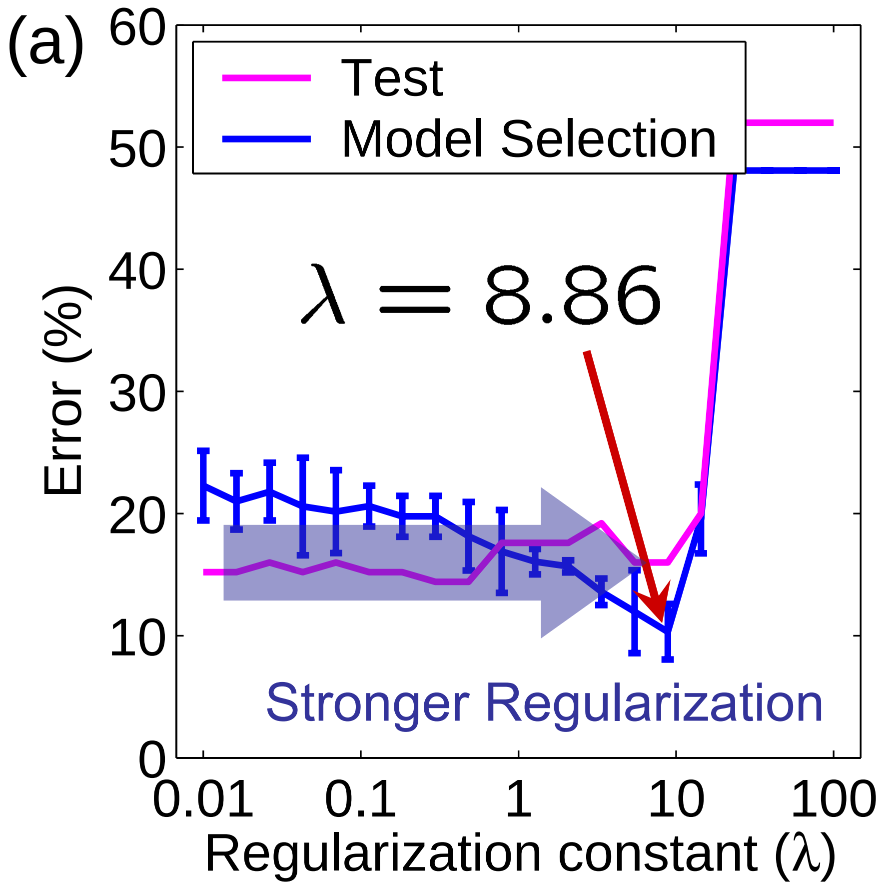
$$W = \frac{1}{2}\left(-\boldsymbol{w}_1\boldsymbol{w}_1^\top + \boldsymbol{w}_2\boldsymbol{w}_2^\top\right)$$

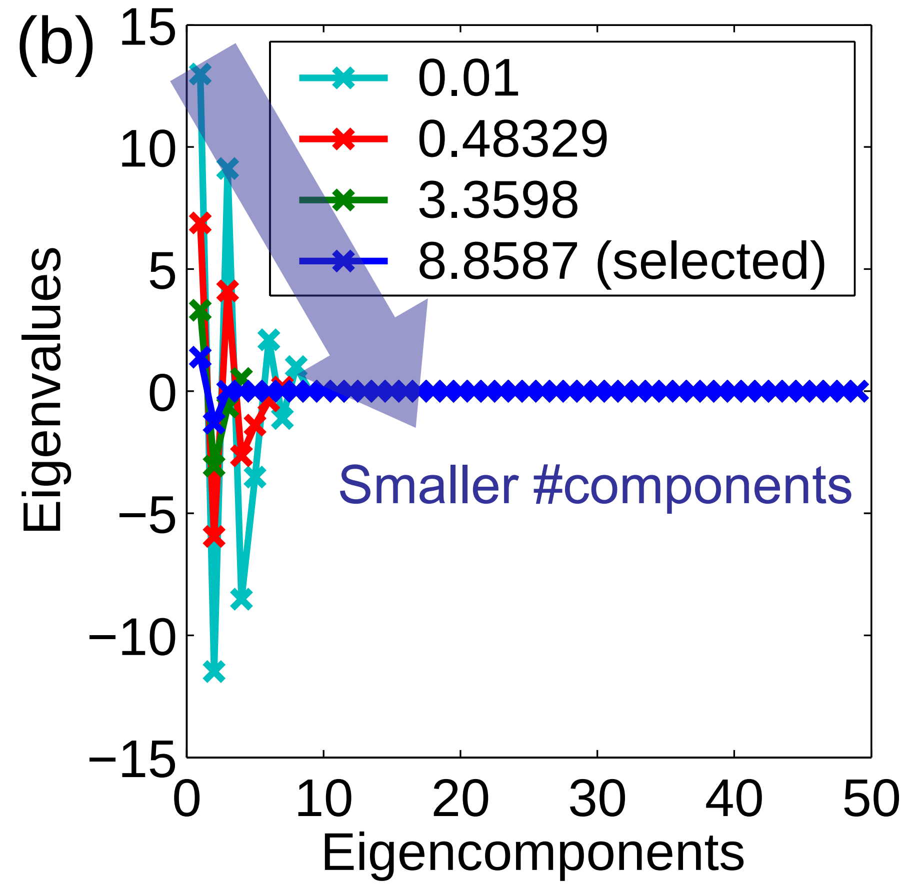
Results: Classification Errors



- Low-ranked (ℓ_1-regularized) solution performs better.
- Fixed rank performs suboptimal.

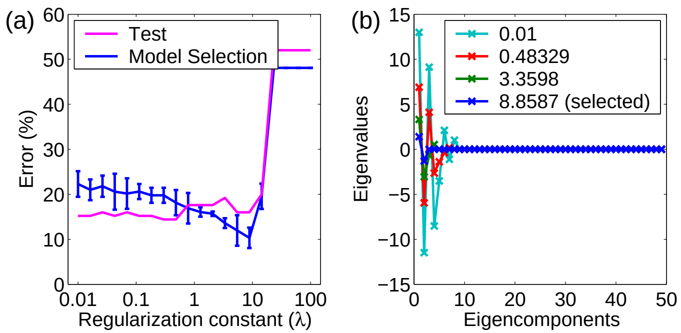
Extracted Features (1/2) Model Selection and Eigenvalues





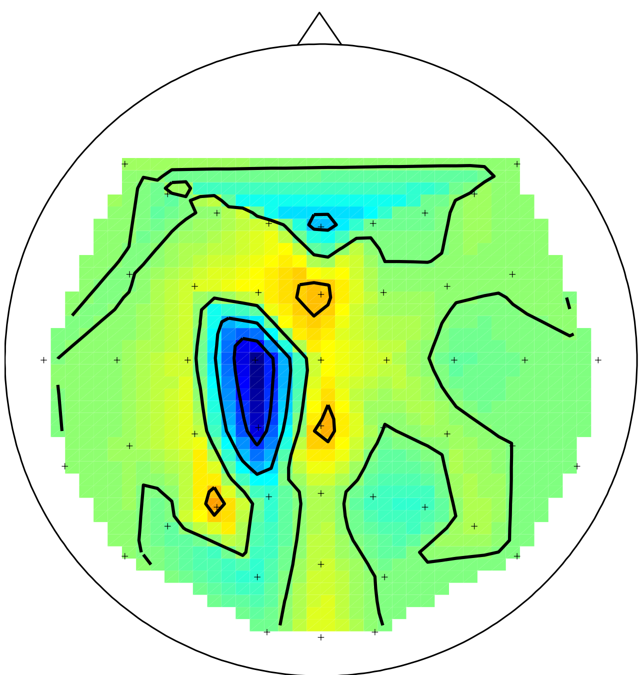
Extracted Features (2/2)

Eigenvectors

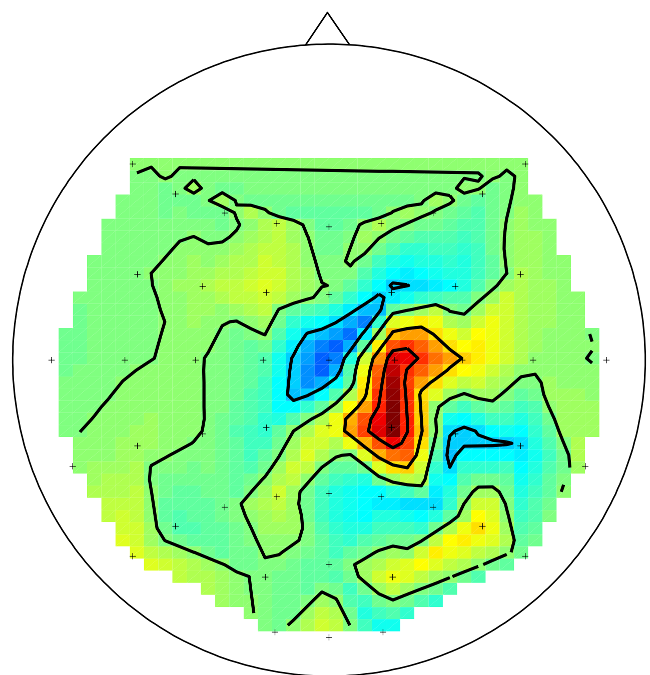


$$W = U \Lambda U^\top$$

U(:,1) $(\lambda_1 = -1.31)$



U(:,2) $(\lambda_1 = 1.40)$



Works on Spectral ℓ_1 (Trace-norm) Regularization

- Prior work by Fazel, Hindi, and Boyd (2001)
- Related work by Abernethy et al. (2006)

	MMMF [Srebro et al. 05]	MTFL [Argyriou et al. 07]	Uncovering Shared Structure [Amit et al. 07]	Classifying Matrices [this talk]
Application	Matrix Factorization	Multi-ouput Regression	Multi-class Classification	Matrix Classification
Loss Function	Hinge-loss	Quad-loss	Hinge-loss	Logit-loss
Input	Scalar	Vector	Vector	Matrix
Output	Matrix	Vector	Vector	Scalar
Optimiza-tion	SDP	Iterative	Primal Gradient	Dual Interior-point

Summary

- Proposed the Matrix Classifier that factorizes using the Spectral ℓ_1-regularization.
  - Single convex optimization problem.
  - Dual formulation and Linear Matrix Inequality for efficient optimization.
  - Sparseness: interpretable solution.
- Applied to motor-imagery EEG classification
  - No distinction between feature extraction step and classification step.
  - Found physiologically relevant features.
  - Application to other problems are in progress.

References

- Koles (1991) "The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG". *Electroencephalogr. Clin. Neurophysiol*., **79**.
- Ramoser et al. (2000) "Optimal spatial filtering of single trial EEG during imagined hand movement". *IEEE Trans. Rehab. Eng.,* **8**(4).
- Fazel et al. (2001). "A rank minimization heuristic with application to minimum order system approximation". *Proc. American Control Conference*.
- Boyd & Vandenberghe (2004). *Convex optimization*. CUP.
- Srebro et al. (2005) "Maximum margin matrix factorization". *Advances in NIPS.* **17**.
- Abernethy et al. (2006) , "Low-rank matrix factorization with attributes"*. Technical report Ecole des Mines de Paris.* N24/06/MM*.*
- Blankertz et al. (2006) "The Berlin Brain-Computer Interface: EEG-based communication without subject training". *IEEE Trans. Neural Sys. Rehab. Eng*. **14**(2).
- Argyriou et al. (2007) "Multi-Task Feature Learning". *Advances in NIPS.* **19**.
- Tomioka et al. (2007) "Logistic regression for single trial EEG classification". *Advances in NIPS.* **19**.
- Amit et al. (2007) "Uncovering Shared Structures in Multiclass Classification". *Proc. ICML.*

Thank you very much!

Derivation of the Dual Problem

(P)

$$\min_{x} f(x) + g(x)$$

Equivalent $\Longleftrightarrow$

(P')

$$\min_{x,y} \quad f(x) + g(y)$$

$$\text{s.t.} \quad x = y$$

$\Downarrow$ Dual

(D)

$$\bar{p}\,(\text{constant})$$

$\Downarrow$ Dual

(D')

$$\min_{\alpha} \; f^*(-\alpha) + g^*(\alpha)$$

Derivation of the Dual

Logistic loss

ℓ_1-norm

(P) $$\min_{\substack{W\in\mathbb{R}^{R\times C}, b\in\mathbb{R},\\ \boldsymbol{z}\in\mathbb{R}^n}} \frac{1}{n}\sum_{i=1}^{n}\ell_{LR}(z_i) + \frac{\lambda}{n}\|W\|_1,$$

s.t. $$y_i\left(\mathrm{Tr}\left[W^\top X_i\right]+b\right) = z_i$$
$$(i=1,\ldots,n),$$

Dual logistic loss

ℓ_∞-norm

(D) $$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \sum_{i=1}^{n}\ell^*_{\mathrm{LR}}(\alpha_i)\Big|_{(0\leq\alpha_i\leq 1)}$$

$$\sum_{i=1}^{n}\alpha_i y_i = 0, \quad \left\|\sum_{i=1^n}\alpha_i y_i X_i\right\|_\infty \leq \lambda,$$

Interpreting the dual variable

$$p_i = \begin{cases} 1 - \alpha_i & (y_i = +1) \\ \alpha_i & (y_i = -1) \end{cases} \quad (i = 1, \ldots, n)$$

$$\text{(D)} \quad \max_{0 \le \boldsymbol{p} \le 1} \quad \sum_{i=1}^{n} H_2(p_i)$$

$$\text{s.t.} \quad \sum_{i=1}^{n} (y_i - \mathbb{E}[y_i | p_i]) = 0,$$

$$\left\| \sum_{i=1}^{n} (y_i - \mathbb{E}[y_i | p_i])\, X_i \right\|_{\infty} \le 2\lambda,$$

Experimental setup

- Offline analysis of 162 datasets from 29 healthy subjects recorded in the Berlin Brain Computer Interface (BBCI) project ([Blankertz et al., 2006], www.bbci.de).
- Binary classification of all the combinations of left hand (L), right hand (R), and foot (F) imaginary movement.
- Multi-channel EEG (32, 64, or 128ch) recordings (70-600 trials in a dataset).
- Band-pass filter 7-30Hz.

Conventional Methods

- CSP (Koles, 1991; Ramoser, 2000) Dimensionality reduction/ demixing technique using label information:

$$\Sigma^{(+)}\boldsymbol{w}_c = \lambda_c \Sigma^{(-)}\boldsymbol{w}_c \quad (c = 1, \ldots, C)$$

$$\Sigma^{(\pm)} = \langle X \rangle_{\pm}$$

$$f(X) = \sum_{c=1}^{C^*} \beta_c \log\left[\boldsymbol{w}_c^{\top} X \boldsymbol{w}_c\right] + \beta_0$$

$$(C^* < C)$$

Conventional Methods

- LR (L2) – Logistic regression with L2-regularization (Frobenius norm)

$$\Omega(W) = \frac{1}{2}\mathrm{Tr}\left[W^\top W\right]$$

- LR (rank=2) – Rank=2 approximated logistic regression

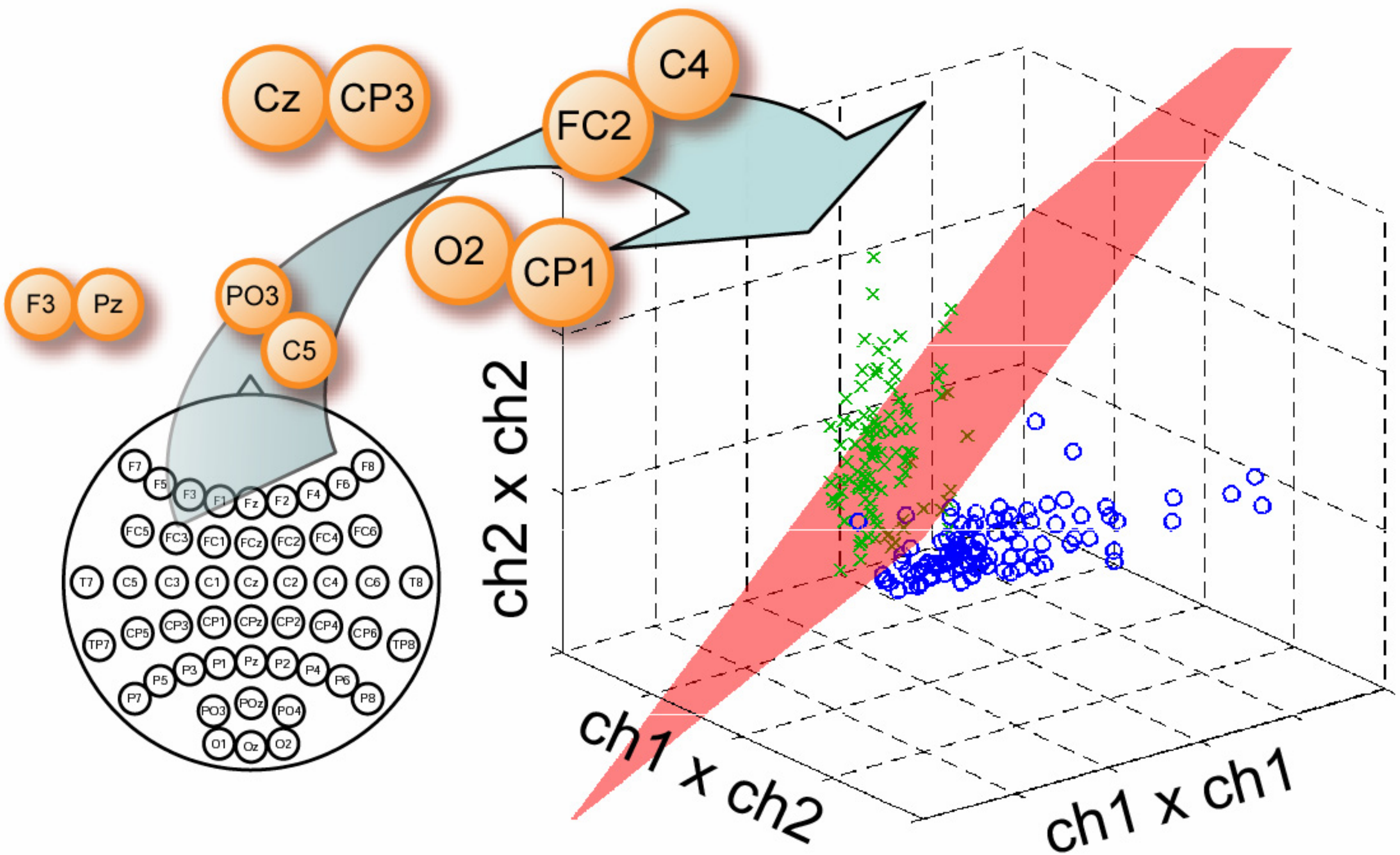
$$W = \frac{1}{2}\left(-\boldsymbol{w}_1\boldsymbol{w}_1^\top + \boldsymbol{w}_2\boldsymbol{w}_2^\top\right)$$

(Tomioka et al., NIPS*2006)

The Discriminative Model

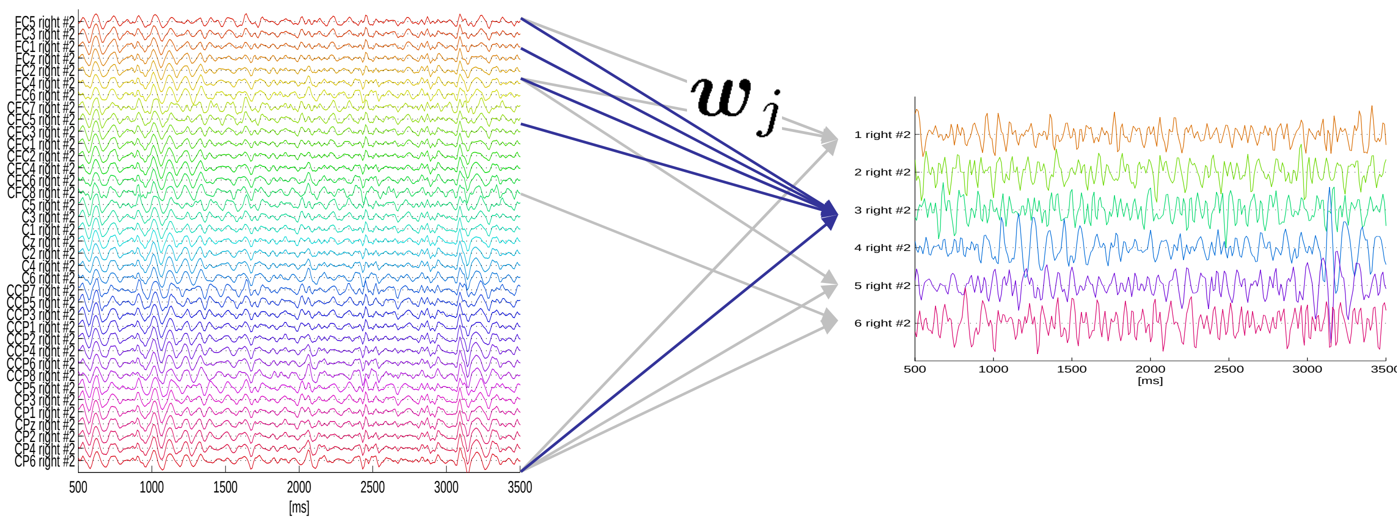
$$f(S; W, b) = \mathrm{Tr}\left[W S S^{\top}\right] + b$$

$$W \in \mathbb{S}^{C}$$
$$b \in \mathbb{R}$$



Appendix: CSP (1/3)

- Common Spatial Pattern (CSP) [Koles, 1991]
  - discriminative dimensionality reduction technique.

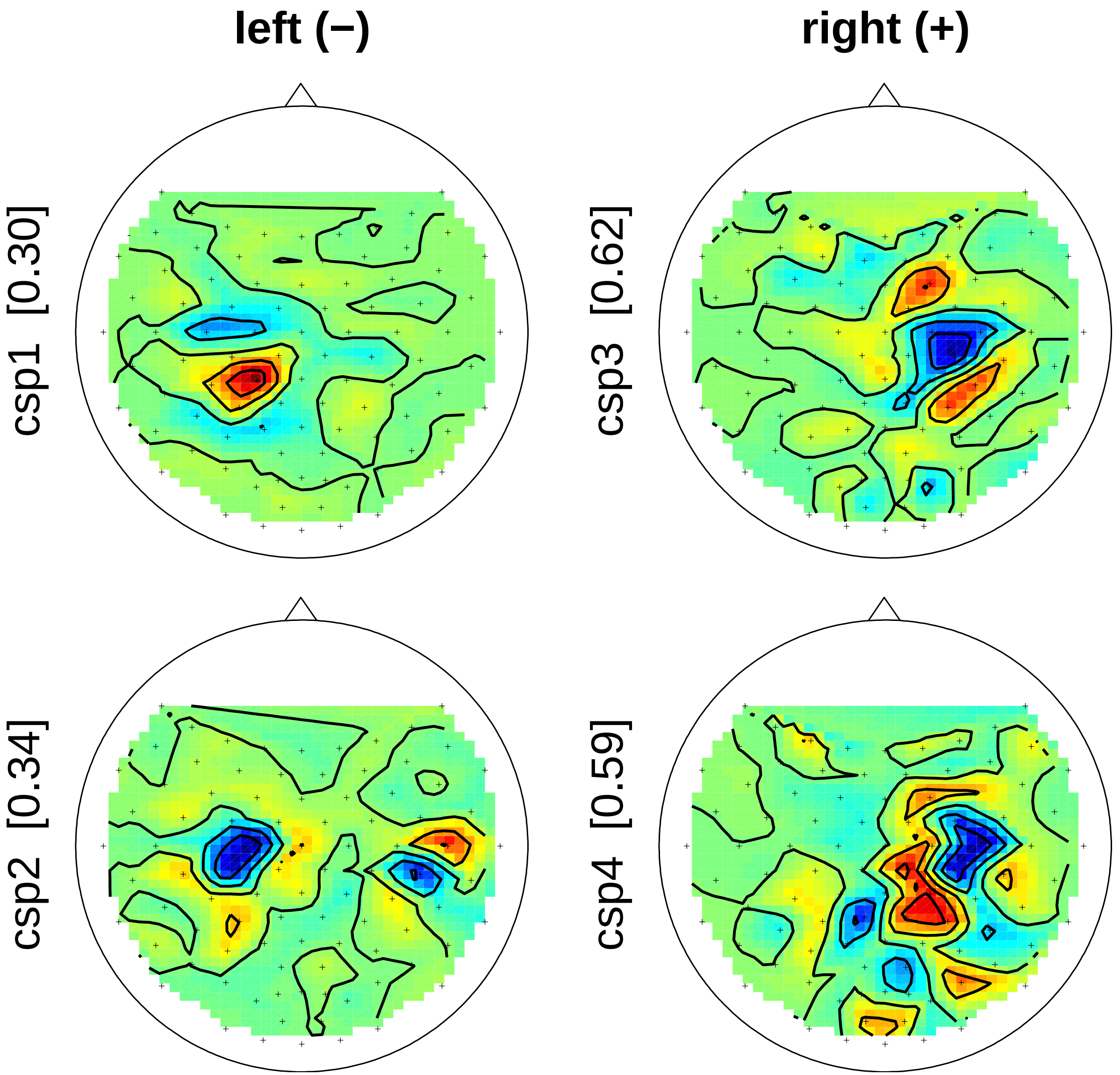


Generalized eigenvalue problem

$$\Sigma^{+} W = \Sigma^{-} W \Lambda$$

Appendix: CSP (2/3)

Example of Spatial Filters



Appendix: CSP (3/3)
CSP filtered time-series